知乎信息起源模型及可信度评估

■ 张婷 齐向华
山西大学经济与管理学院 太原 030006

摘要：[目的/意义]构建知乎信息传播过程PROV数据起源模型和用户可信度评价指标,量化知乎信息的可信度,丰富和完善社会化问答社区平台信息可信度评估的方法。[方法/过程]以知乎为研究对象,从信息传播过程的角度出发引入数据起源概念评估信息的可信度,通过建立知乎的PROV数据起源模型,追溯并记录知乎信息的来源和传播过程,与信息传播过程中涉及到的用户可信度分值相结合,计算出知乎信息的可信度定量结果。[结果/结论]通过对知乎信息可信度的评估,进一步完善信息可信度评估方法,为优化社区信息质量提供新思路。

关键词：数据起源 PROV模型 知乎信息 可信度
分类号：G252
DOI:10.13266/j.issn.0252-3116.2019.09.009

随着互联网的发展,在信息的爆炸式增长和无序的状态下,信息流动的方向发生了根本性变化,用户主动获取信息的目的性增强,获取方式不再局限于通过搜索引擎获取信息,而是通过在线提问和社区讨论的方式从同行、专家处获取经验、信息和知识。知乎作为一个为用户提供彼此分享知识、经验和见解的社会化问答社区网站,截至2017年9月,其个人注册用户总数超过1亿,日活跃用户量达2 600万。人均日访问时长1小时,月浏览量可达180亿次[1]。据中国互联网络信息中心发布的第41次《中国互联网络发展状况统计报告》,截至2017年12月,知乎用户使用率从2016年的7.0%上升到8.8%[2]。但是大量用户涌入导致社区内容变化的不可控严重冲击了知乎对自身社区"高质量"的要求和定位。同时知乎对于热门话题的多维度见解和多角度分析,使其成为舆论发酵地,或是观点二次传播的重要节点,对用户观点探讨的深度和事件发展的方向具有一定的推动作用。用户数量不一定是直接导致知乎内容质量降低的主要原因,信息的高速流动而造成的信息泛滥才是导致内容质量良莠不齐的重要因素之一。高质量的信息内容能够为用户乃至互联网提供高价值的知识资源,因此,如何从大量的知乎信息中识别可信度高的答案成为一个亟待解决的问题,对于知乎问答的信息的可信度研究变得至关重要。

信息可信度一般指信息或信息源被信息接受者信任的程度,在这里解释为用户对信息来源或传播介质的可信赖性以及信息发布者的权威性的感知或评价。目前学者对于知乎和其他问答社区网站的信息可信度评估主要从用户和信息文本内容两个独立的角度以及基于分类特征的信息质量特征提取和信息质量评估影响建模两个方面进行研究[3],缺少从信息产生和传播过程的角度来评估信息可信度的研究。数据起源可以帮助追踪数据的来源,记录数据处理过程中产生的动态信息并评估数据质量以及可信度,使用户清晰地了解到信息的来龙去脉,并根据自己的需求决定是否采用该信息内容。

本文以知乎为研究对象,采用数据起源方法对知乎信息的可信度评估进行研究,从信息传播过程这一角度评估在线问答社区信息可信度,为相关研究提供新的思路。分析知乎信息传播的场景并构建PROV数据起源模型,抓取信息传播过程中所涉及的知乎用户网络行为数据,通过为他们分配信任分数来估计信息提供者的可信度,并基于传播路径计算出知乎信息可信度得分。基于信任得分,用户可以对是否使用该信息做出更明智的决定。

1 研究现状

与传统的基于搜索引擎的问答社区网站不同,知

作者简介:张婷(ORCID:0000-0003-0992-0789),硕士研究生,E-mail:1158229980@qq.con;齐向华,教授。
收稿日期:2018-07-18 修回日期:2018-11-14 本文起止页码:85-94 本文责任编辑:刘远颖

乎引入了社交关系,以互动问答为核心,进一步促进知识的分享和互动,可以称之为真正意义上的社会化问答社区平台。截至目前,国内关于社会化问答社区平台的研究大多集中在平台特点与发展模式以及用户行为研究等方面,而关于平台信息质量和信息可信度的研究成果不多。姜雯、王平、李保珍等学者[4-6]对国内外的网络信息质量评价的相关研究进行梳理和综述。通过对国内外相关文献的阅读和整理,问答社区信息质量评估主要从分类特征的信息质量特征提取和信息质量评估影响建模两个方面进行研究。

1.1 分类特征的信息质量特征提取

基于此方法的相关研究有:J. Jeon[7]对 Naver Q&A 社区中的 1 700 个答案采用最大熵算法,分别从回答者采纳率、专业性以及回答长度、数量、被复制次数等进行分析。E. Agichtein[8]采用随机梯度增强决策树分别对 Yahoo! 上 6 665 个问题和 8 366 个答案进行特征提取。C. Shah 等[9]采用逻辑回归算法对 Yahoo! Quest 中 120 个问题和 600 个答案进行特征分析。Q. Tian 等[10]采用随机森林对 stackoverflow 中 103 793 个问题和 196 145 答案进行特征分析。来社安[11]采用 SVM 算法对百度知道中 12 707 个问题及其 463 114 个答案进行特征提取。王伟[12]对知乎的 20 个话题、200 000个问题、924 266 个回答、158 007 个用户数据采用逻辑回归、支持向量机和随机森林的研究方法构建特征体系。

1.2 信息质量评估影响因素建模

S. Oh 等[13]邀请不同职能的人在选定的评价标准下对 Yahoo! Answers 的有关健康问题的答案的准确性、完整性、相关性、客观性、来源可靠性等相关指标进行评估。国内学者贾佳等[14]采用问卷调查的方式对知乎、百度知道平台的答案质量进行对比和打分。P. Fichman[15]针对 Wikipedia Reference Desk、WikiAnswers、Yahoo! Answers 和 Askville 4 个问答网分别从准确性、完整性、可证实性 3 个方面采用人工打分的方法比较评估了回答的质量, 所得结论为流量高的问答网站与答案的质量不成正比。S. Kim 等[16]运用内容分析的方法,对 Yahoo! Answers 里用户在选择最佳答案时所写的评论进行了分析,结果表明,话题分类不同,其评价模式也有所不同,用户选择最佳答案的主要标准是社会性情感、内容和效用。李晶[17]和孙晓宁[18]均采用结构方程模型方法构建百度知道的信息质量模型和 SQA 系统答案质量模型。曹高辉等[3]以 Quora 和知乎两个平台为研究对象,采用问卷调查的形式获取到用户对社会化问答平台答案信息质量感知,并在此基础上构建出答案质量感知的外部模型。施国良[19]以知乎为研究对象,从答案特征和答题者特征两个角度提出了研究假设,采用内容分析法和回归模型对假设进行检验,并得出答案长度、答案及时性、答题者的影响力均正向影响答案认可度以及答题者特征比答案特征对答案认可度的影响更大的研究结果。

综上所述,分类特征的信息质量特征提取是基于庞大的数据量的研究方法,也是大数据时代计算机自动化评价的趋势,具有效率高、分析力强的特点,但存在难以处理用户情感态度的主观性问题。而信息质量评估影响因素建模的研究方法弥补了机器难以理解主观问题的特点,通过对不同人群给出的评价而得出具体的影响因素,从而建立信息质量模型。但此方法只能反映出小部分数据的情况,且评估结果受评估者自身信息素养水平、问题意识以及专业性等主观因素影响。

2 数据起源模型

PROV 是万维网联盟起源工作组(W3C Provenance Incubator Group)定义的以 PROV 概念数据模型(PROV-DM)为核心的 12 个文档组成的溯源标准文档集合,旨在追溯网络所需要的源数据,描述了“实体”的产生、“活动”的产生以及“代理”的责任等动态信息,以实现对数据的溯源和规范化表达。核心架构包括实体、活动、代理三类,三者之间的 7 个关系和衍生关系,根据活动情景的特点和需要,对实体、活动和代理之间的关系提供了进一步的描述方式,具体的扩展条款见图 1。

实体与实体之间产生变化通过属性 prov:wasDerivedFrom 来表示,为了更详细地描述不同情景中实体和实体之间的关系,其进一步提供了一般属性,如衍生类的 3 个子属性:prov:wasQuotedFrom 表示引用较大的书本、图集或网络来创建新的实体,新实体重复部分或全部原实体[20];prov: wasRevisionOf 表示派生实体包含来自原始实体的大量内容,例如两本不同版本的图书;prov:hadPrimarySource 表示新实体的内容主要来源于某主题或原实体里的经验或知识。prov:specializationOf 和 prov:alternateOf 从抽象级别上进行深度的关系描述,前者将具有详细内容或更具说明性的实体连接到简单概括的通用实体上,后者表示链接对同一事物或主题的其他属性补充的实体。属性 prov:wasInfluencedBy 关联了受影响或影响其特性的实体、活动或

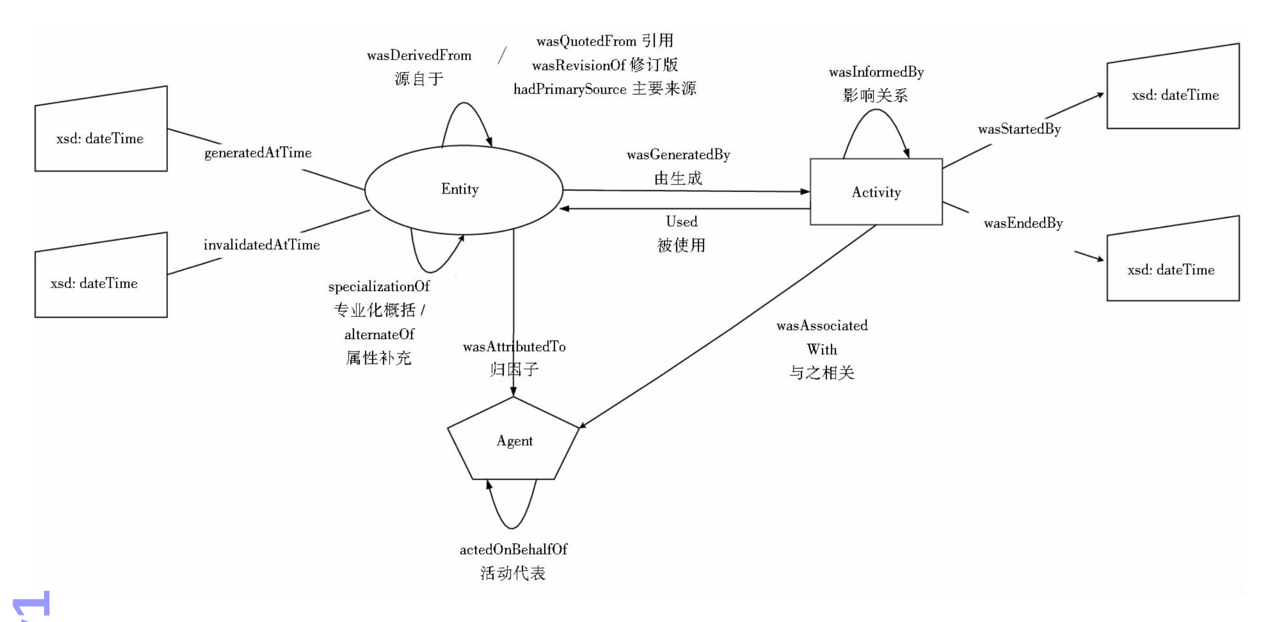


图1　PROV 核心架构

代理。而实体的生成时间和失效时间分别用属性 prov:generatedAtTime 和 prov:invalidatedAtTime 表示。活动和活动之间在时间上存在先后顺序，用属性 prov:wasInformedBy 表示；其开始时间和结束时间分别用属性 prov:wasStartedBy 和 prov:wasEndedBy 来描述。代理之间的关系使用属性 prov:actedOnBehalfOf 表示；实体和活动之间存在两种关系，分别是 prov:used 和 prov:wasGeneratedBy，表示实体被活动所使用和实体由活动产生；代理与实体之间的归属关系用属性 prov:wasAttributedTo 表达；代理与活动的相关关系用属性 prov:wasAssociatedWith 表达。

3　知乎信息可信度评估分析

知乎作为一个基于社会网络关系并引入维基百科社区精神的社会化问答社区平台，其提供的是一个产生、分享和传播知识的工具和框架，而内容由用户发布，用户将自身的隐性知识显性化，通过用户节点将知识进行分享和传播。而知乎信息的传播模式主要是通过人和信息流（包括话题、问题和答案），将人与信息以及信息与信息连接起来，抽象为模型[21]，见图2。

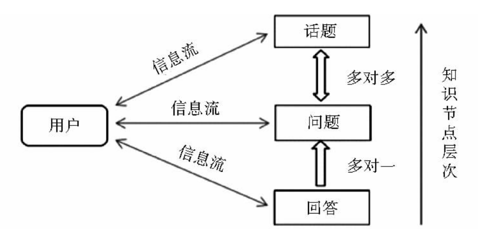


图2　知乎信息传播抽象化模型

知乎的核心内容来自问答框架，基于话题，以问题为中心，所有回答必须围绕问题本身。因此知乎根据用户的回答采取“赞同—反对”机制让用户发起评判，知乎用户可以通过赞同、感谢、收藏、评论、关注、私信、打赏、分享、反对等方式参与互动，再根据评判后的评分情况把认同度更高的回答放在更靠前的位置，并且问题的每个回答下都显示着所获得的点赞数量，以此让用户对该答案的质量有直观的判断。针对不良内容，知乎通过“系统判定辅助＋人工审核”进行核实处理，采用名为“瓦力”和“悟空”的 AI 反垃圾系统对错误内容进行实时筛查，或是对无效回答、失实内容、垃圾广告导流等行为进行识别和处理。在此基础上，对于算法无法处理的内容则通过人工审核该回答是否解释了该问题，并对答非所问和违反规范的内容进行折叠，通常在回答列表的底部集中显示。

以上是基于用户、社区管理人员以及算法对回答内容上的直观判断，若一个回答的内容源自于一个权威的组织，其可信度较高，若源于一个普通用户，其可信度就该怎么判断呢？若该用户根据其他人的评论对自己的回答进行修改和补充，那这个回答的可信度是否可以被视为较高呢？若该用户回答的内容是经过多次转载，并且被多次转载的内容通常是经过他人判断，且包含着一些认知在其中，那这个回答的可信度高吗？因此，记录信息的来源、变化以及传播过程将变得有价值。本文利用 PROV 起源模型着眼于活动对资源的影响[22]这一特点，通过描述影响知乎信息可信度的所有活动创建内容和用户的信息链，以追溯信息的来源和内容变化的不同版本。又由于答案信息的产生是由信

息发布者根据自己对答案的理解，将自己的隐性知识显性化的过程，信息发布者与答案信息的质量具有直接关系。所以影响内容型社区“高质量”的两大重要因素为信息传播过程和用户关系。根据所确定的影响因素，选择相应的评估方法：

（1）针对于信息传播过程，抓取所选的知乎信息传播场景的起源信息，对其进行存储和校验，通过建立 PROV 模型来确定信息传播路径。

（2）对信息传播过程中涉及到的用户进行可信度评估，根据知乎平台的用户特点和用户行为特征来确定用户可信度评估指标，并计算得出定量结果。

（3）将传播路径与用户可信度值相结合，计算出知乎信息可信度值。

3.1 知乎信息传播路径分析

3.1.1 知乎信息传播场景分析　根据以上分析可知知乎中有以下情景（见图 3）：选取知乎圆桌中 Topic1 的问题，用户 user_A 于 2018 年 2 月 5 日 12:18 时给予了回答 answer_a，同时，用户 user_B 引用自己曾发布过的一篇文章给予了回答 answer_b，并在文章的结尾处附有引用文章的链接。用户 user_C 和用户 user_D 给 answer_b 给予了评价 comment_a 和 comment_b 且均被选为精选评论，同时用户 user_D 对 answer_b 点赞。在 2018 年 2 月 6 日，用户 user_E 给出了与之前用户观点不同的回答 answer_c，且内容摘录于相关专业的论文。用户 user_F 给出了评论 comment_c，且用户 user_G 根据问题和 comment_c 给出了 answer_d。用户 user_A 根据 comment_a 和 comment_b 对 answer_a 进行编辑修改，给出了回答 answer_a1，且在 2018 年 2 月 11 日根据 answer_d 再次编辑修改答案，给出了回答 answer_a2。

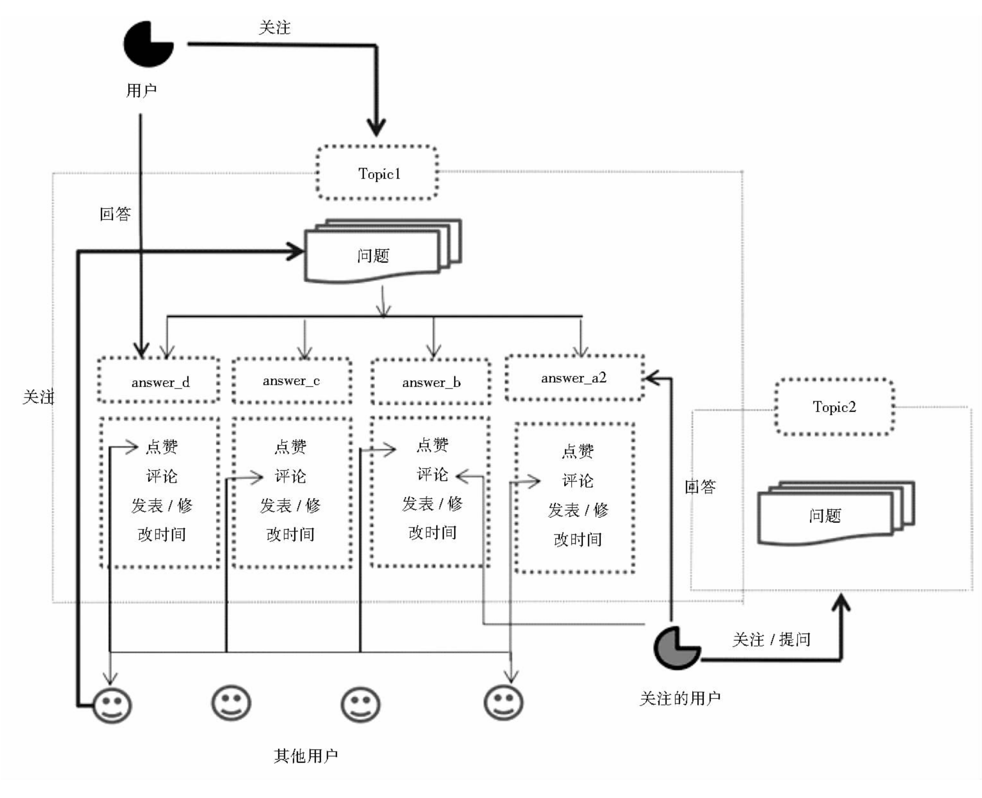


图 3　知乎信息传播场景

3.1.2 知乎的 PROV 数据起源模型　根据数据起源模型的描述规则可知，起源资源有对应的唯一的 URI 进行标识，且命名空间由 URI 标识，本文采用示例命名空间 http://example.org/对资源进行标识，因此，在知乎信息传播的过程中涉及的实体、活动以及代理都用简写的命名空间前缀表示。

（1）实体（entity）。PROV 起源模型中是指客观存在的、概念上的各种类型的事物。对于知乎信息传播的过程而言，实体主要包括答案、文章、评论、转载信息链接等。

（2）活动（activity）。PROV 起源模型中用来描述实体如何维持现状以及通过改变实体的属性从而产生新的实体的过程，诸如行动和过程等。知乎信息传播过程中的活动主要包括答案发布、编辑、评论、点赞等活动。

（3）代理（agent）。PROV 起源模型中对活动、实体

和角色承担负责。它可以是个人、组织以及软件等无生命的物体。一个活动可能与若干个代理关联，且其中涉及的相关实体也与代理存在着因果关系。在知乎信息传播过程中的代理是指知乎的用户。

(4)衍生和修订(derivation and revision)。PROV起源模型中描述的是实体与实体间的关系，一个实体推导出另一个实体，后者的内容、属性以及存在源于前者。在知乎中，用户根据之前用户给出的答案以及相关评论结合自己的经验或专业给出新的回答。修订属于特殊的推导，随着时间的变化，用户对之前的回答进行多次编辑修订，PROV起源模型中，每次编辑产生的新版本都会被视为新的实体。

(5)时间(time)。PROV起源模型中是用来描述活动的生成与结束的时间，也是对活动和实体之间的关系的进一步说明和描述。在知乎信息传播的过程中主要体现在回答内容发布的开始时间和编辑修改的结束时间，时间较近的回答可能是修改过或相对较新的想法，其可信度相对更高。

知乎信息传播过程涉及的对象描述如表1所示：

表1 知乎信息传播过程涉及的对象描述

类型	对象	描述
实体	answer	答案
	article	文章
	commentInfo	评论
	repostLink	转载信息链接
活动	publish	答案发布
	compile	编辑
	comment	评论
	approve	点赞
代理	user	知乎用户
衍生和修订	prov: wasDerivedFrom	衍生
	prov: wasQuotedFrom	引用
	prov: wasRevisionOf	修订
	prov: hadPrimarySource	主要来源
时间	prov:generatedAtTime	实体生成时间
	prov:startedAtTime	活动开始时间
	prov:endedAtTime	活动结束时间

根据知乎信息传播情景，运用PROV起源模型，可以得到完整数据溯源描述，见图4。分别用W3C的数据起源模型描述规范中的椭圆形表示实体，矩形表示活动，五边形符号表示代理。代理的可信度用prov:userTrust表示。箭头和无框的文本用来表示和说明三者之间的使用和产生的关系，箭头方向由未来指向过去。代理是为活动或实体负责的，箭头方向由活动指向代理，代理与活动的连接采用wasAssociatedWith表示，代理与实体的连接采用wasAttributedTo表示。

3.2 用户的可信度评估

知乎的核心用户为该平台提供了少有的经验以及独特的解决问题的思路等高质量内容，是提高知乎核心竞争力的重要力量。因此，知乎要求用户提供完整的个人信息，并采取实名制认证其身份，通过统计其在使用期间获得的赞同、感谢、收藏次数等个人成就来增加其回答内容的可信度。因用户与用户之间存在社交属性，用户发布答案、想法和评论的频率越高，与其他用户信息交流的速率就越快，其个人的被信任程度也会随活跃度的增高而增高。同时，在知乎用户的关系网络中，拥有较多关注者的用户，其回答的内容可以被更多的人阅读、赞同和评论，以提高自身交互能力的程度。用户与关注者的交互能力强，其可信度随着影响力的增大而增高。同时用户也可以关注其他用户，能及时获取他人的有用信息并丰富和提高自身的知识储备和回答内容的质量。因此，本文将从用户信息完整度、用户认证和成就、用户活跃度和用户交际广度[23]4个方面对用户的可信度进行评估。

3.2.1 用户信息完整度 该指标在知乎社区里具体包括头像、地址、教育信息、职业信息、个人简介。用户资料完整度越高，大众对其信任程度会更高。用户基本资料用向量 $Y=(x_1,x_2,\cdots,x_n)$ 表示，计算用户信息完整度的公式如下：

$$x(i)=\begin{cases}0, & x(i)\text{为无效信息}\\1, & x(i)\text{为有效信息}\end{cases} \quad \text{式 (1)}$$

$$UI(u)=\frac{1}{n}\sum_{i=1}^{n}x(i) \quad \text{式(2)}$$

其中，公式(1)表示第i条是否存在有效信息。公式(2)UI(u)表示用户资料完整度函数。

3.2.2 用户认证和成就 知乎平台对用户信息的真实性具有一定的要求，这也是评估用户可信度的重要条件之一。知乎提出了个人与机构的蓝色认证机制。但随着大量用户的涌入和回答数量的增大，为了让用户更加高效地识别各领域的专家与内容，知乎对在特定话题下创作了大量专业内容的用户进行了“优秀回答者”标识，即橙色标识的认证[24]，优秀回答者均由系统通过话题权重计算而来，提高自己在知乎某个领域下的权重只有一个方法，就是在该领域下发布高质量的回答。相比蓝色认证，橙色认证更能进一步代表用户的可信度。成就是指其他用户对该用户的赞同、感谢、收藏的情况，具体公式如下：

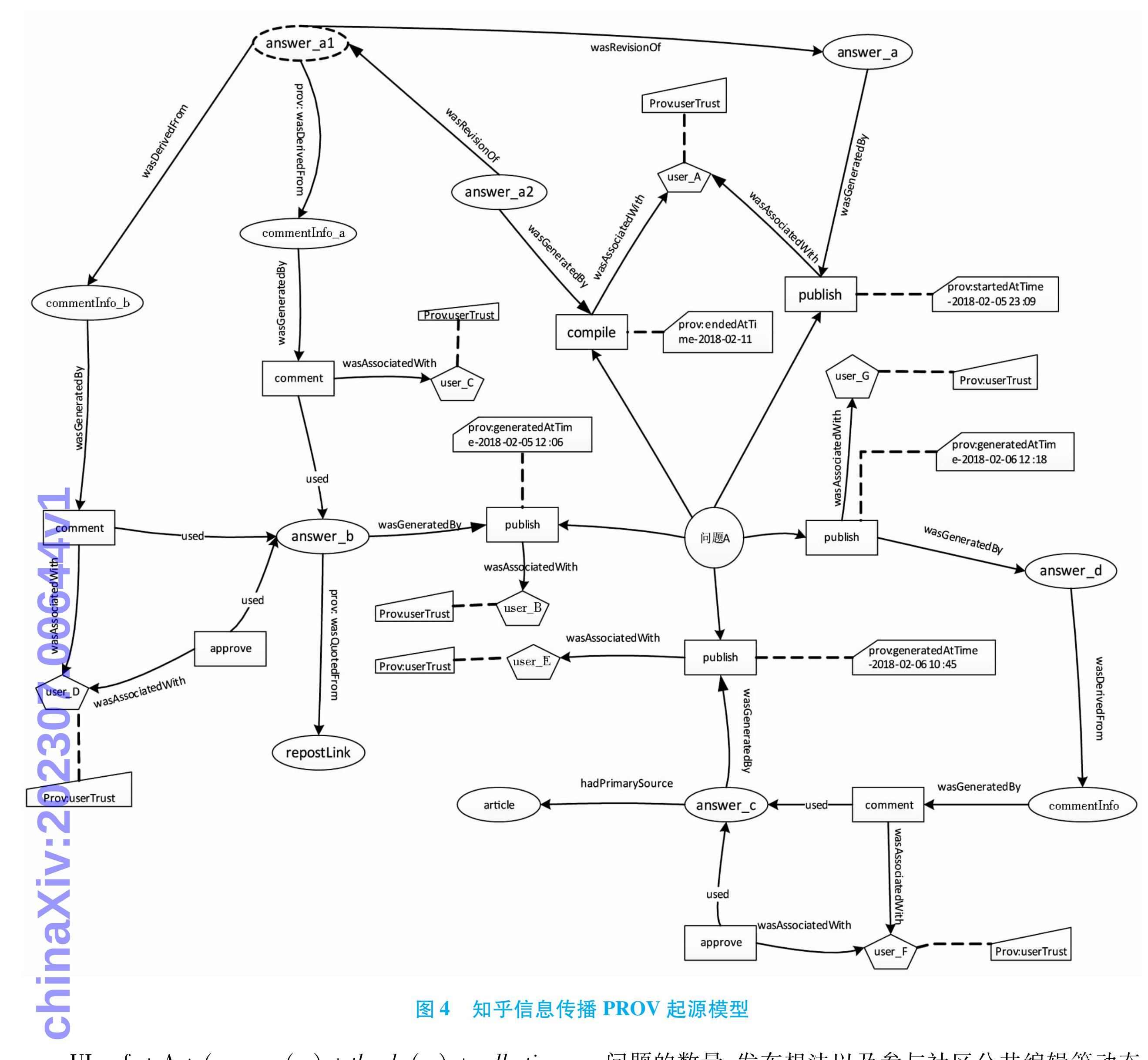


图 4　知乎信息传播 PROV 起源模型

$$UL = f_j + A * (approve(n) + thanks(n) + collection(n)) \qquad \text{式(3)}$$

不同认证类型的用户拥有不同等级用户的可信度评分数，如公式(3)所示(所涉及的用户认证等级和成就权重系数见表 2)，其中 fj 为类型 j 的用户认证等级评分。n 为其他用户对用户的所有回答的赞同、感谢和收藏次数。A 为针对用户成就对可信度的影响力即成就权重系数，采用乘积标度方法设定的信任度权重系数。

表 2　用户认证等级和成就权重系数

j	f_j	n	A
未认证用户	0.19	赞同次数	0.24
蓝色认证	0.35	感谢次数	0.44
橙色认证	0.46	收藏次数	0.32

3.2.3　用户活跃度　指用户在一定时间内点赞、回答问题的数量，发布想法以及参与社区公共编辑等动态次数。公共编辑是用户主动参与改善和优化知乎公共内容的行为，因此用户活跃度越高，表明其社区参与度越高，可信度就相对越高。公式如下：

$$UC = \frac{\sum_{i=1}^{T}(publish(i) + approve(i) + idea(i))}{T} + B * ed \qquad \text{式(4)}$$

公式(4)中 publish(i)、approve(i)和 idea(i)分别指知乎用户在第 i 天回答的问题数，赞同数和发布的想法数，T 为设定的某一时间段。ed 是指知乎用户参与公共编辑的次数，其设定的信任度权重系数 B 为 0.44。

3.2.4　用户交际广度　关注用户的粉丝越多，其对信息扩散的可能性越大，影响力越大。若粉丝的知名度越高，说明该用户的可信度更高。因此粉丝数比用户

关注的人数更影响用户的交际广度。

$$UD = C * lev(fans) + D * lev(followers) \quad 式(5)$$

由公式(5)可知,fans 指关注用户的人数;followers 指其关注的用户数;Lev(n):代表用户所关注或关注者的认证等级,n 越大,等级越高;C、D 作为信任度权重系数分别为 0.32 和 0.24。

由以上 4 个用户指标分析,计算可得用户的可信度(User Trust),公式如下:

$$UT = \frac{UI + UL + UC + UD}{4} \quad 式(6)$$

3.3 知乎信息可信度计算

通过 PROV 数据起源模型对知乎信息的来源以及在传播过程中内容的变化情况进行追踪,可清晰地得到信息传播路径,并与传播路径上节点的用户可信度相结合,计算得出知乎信息的可信度。

$$Trust(answer) = \frac{\sum_{i=1}^{n} userTrust(i)}{n} \quad 式(7)$$

公式(7)中,n 为针对一条回答的传播路径上涉及的用户数。Trust(answer)为知乎信息可信度。

4 实验设计与分析

根据本文先前描述的知乎信息传播的场景,收集相应的起源信息,根据起源模型要求的 RDF 语句规范进行描述,对其进行存储和校验,最终建立 PROV 起源模型。根据之前设定的指标对传播过程中涉及的用户的可信度进行分析,结合基于数据起源的知乎信息传播过程,计算出知乎信息的可信度评估值。

4.1 起源信息预处理

通过收集记录知乎信息传播路径的信息,对传播路径节点上涉及的 7 个测评用户采用八爪鱼采集器来采集其个人信息和行为动态,采集的内容是基于以上所确定的 4 个用于评估用户可信度指标和能够明确反映或间接挖掘出用户可信度的信息。其中回答问题数、赞同数、发布想法数采集了用户一个月内的动态,具体如表 3 所示:

表 3 用户信息指标实际值

用户 ID	点赞次数	感谢次数	收藏次数	公共编辑次数	关注人数	关注者人数	回答问题数	赞同数	发布想法数
Mantou0322	776	171	1 079	10	17	302	0	0	0
ding-xiang-yi-sheng	1 605 625	297 112	1 012 938	0	229	1 179 225	5	7	5
xu-zhao-qing	25 740	3 989	3 472	112	120	3 060	26	1	5
ceng-jin-cheng	47	15	5	17	95	9	1	1	0
ke-xue-shi-jie-za-zhi	63 002	9 574	29 109	0	44	60 927	6	12	1
greatfire	427	65	18	61	676	26	0	15	0
Long-ya-57-84	881 402	99 820	135 748	40	68	140 789	37	3	1

资源描述框架(Resource Description Framework, RDF)是用来描述网络资源的 W3C 标准,RDF 事实上已成为 PROV 模型的标准描述方式[25],本文对收集到的信息采用 RDF 三元组的形式进行描述,选取部分 RDF 描述如下所示:

```
ex:publish prov:wasAssociatedWith ex:user_A
ex:compile prov:wasAssociatedWith ex:user_A
ex:comment prov:wasAssociatedWith ex:user_C
ex:approve prov:wasAssociatedWith ex:user_D
ex:comment prov:used ex:answer_b
ex:approve prov:used ex:answer_b
ex:approve prov:used ex:answer_c
ex:answer_a2 prov:wasGeneratedBy ex:compile
ex:answer_b prov:wasGeneratedBy ex:publish
ex:commentInfo_a prov:wasGeneratedBy ex:comment
ex:commentInfo_c prov:wasGeneratedBy ex:comment
ex:answer_a a prov:Entity
ex:repostLink a prov:Entity
ex:commentInfo_a a prov:Entity
Ex:answer_b a prov:Entity
ex:user_A a prov:Agent,"Person"^^xsdatring
prow:userTrust"0.521"^^xsd:double
ex:user_B a prov:Agent,"Organization"^^xsdatring
prow:userTrust"0.863"^^xsd:double
ex:user_C a prov:Agent,"Person"^^xsdatring
prow:userTrust"0.656"^^xsd:double
ex:user_D a prov:Agent,"Person"^^xsdatring
prow:userTrust"0.355"^^xsd:double
ex:user_E a prov:Agent,"Organization"^^xsdatring
prow:userTrust"0.355"^^xsd:double
ex:user_F a prov:Agent,"Person"^^xsdatring
prow:userTrust"0.355"^^xsd:double
prov:startedAtTime "2018-02-05T23:09:00"^^xsd:dateTime.
ex:publish a prov:Activity;
prov:endedAtTime "2018-02-05"^^xsd:dateTime.
ex:compile a prov:Activity;
```

prov:generatedAtTime "2018-02-05T12:06:00"^^xsd:dateTime.
ex:publish a prov:Activity;

4.2 用户可信度计算

从该社会化问答平台中采集有关信息发布者的相关数据如表 3 所示,为了减少量纲不同以及数值数量级间的悬殊差别对用户可信度计算结果所带来负面影响,本文对原始数据即指标实际值采用标准化方法进行无量纲化预处理,再进行可信度相关指标的计算,公式如下:

$$y_i = \frac{x_i - \bar{x}}{s} \quad 式(8)$$

公式(8)中 x_i表示指标实际值,$\bar{x}$ 为指标值均值,s 为指标的标准差,y_i 表示指标评价值。经过初步数据处理,以及对四个评估指标的计算,分别得出评估指标的定量结果和用户可信度值,如表 4 所示:

表 4 用户可信度相关指标分值

用户编号	用户 ID	信息完整度	认证和成就	活跃度	交际广度	用户可信度
user_A	Mantou0322	0.8	0.276	0.179	0.135	0.348
user_B	ding-xiang-yi-sheng	1	3.541	0.638	1.221	1.6
user_C	xu-zhao-qing	0.6	0.303	2.344	0.244	0.873
user_D	ceng-jin-cheng	0.8	0.507	0.321	0.216	0.461
user_E	ke-xue-shi-jie-za-zhi	1	0.521	0.704	0.207	0.608
user_F	greatfire	0.8	0.275	1.228	0.823	0.782
user_G	Long-ya-57-84	1	1.119	0.692	0.291	0.776

4.3 知乎信息评估结果与分析

根据图 4,知乎信息传播 PROV 起源模型可以对知乎信息 answer_a2 进行追踪,用户 user_A 针对于用户 user_C、user_D 的评论对已有的评论 answer_a1 在 2018 年 2 月 11 日进行了重新编辑从而生成 answer_a2,由此得到该回答生成的整个过程。因此 answer_a2 的可信度计算公式为:

$$Trust(answer_a2) = \frac{\sum_{i=1}^{3} userTrust}{3} \quad 式(9)$$

通过对用户信息的完整度、个人认证和成就、活跃度、交际广度 4 个指标计算出用户的可信度值,结合 PROV 数据起源模型,通过公式计算得出知乎信息可信度结果从高到低排序如表 5 所示:

表 5 知乎信息可信度计算结果

信息编号	路径涉及用户	发布时间	可信度值
answe_b	user_B	2018 - 2 - 5 10:48	1.6
answe_d	user_G、user_F	2018 - 2 - 6 12:06	0.779
answe_c	user_E	2018 - 2 - 6 12:18	0.608
answe_a2	user_A、user_C、user_D	2018 - 2 - 11	0.56
answe_a	user_A	2018 - 2 - 5 23:09	0.348

从计算结果的排序中可以看出信息内容的可信度与发布者在知乎平台上的认证以及成就具有正向关系。作为具有权威性和专业性的组织 user_B,其用户可信度值最高,发布的信息内容 answer_b 是参考其组织内部发布的具有权威性的文章,从计算结果可以看出 answer_b 的可信度最高,与实际情况相符,并且与知乎基于算法和"赞同—反对"机制所给出的排序相符。虽然信息具有时效性和滞后性,但对于此次实验样本而言,相比较用户的可信度,时间对信息可信度的影响较小,若时间作用到用户身上,使用户的个人认知变化随时间的变化以及其他用户的影响而得到持续的优化和改进,则时间对信息的可信度有一定的正向影响。answer_a 的发布者 user_A,既没有得到知乎平台的个人或组织认证,其活跃度、交际广度也相对较低,其发布的内容 answer_a 的可信度比其他用户发布的内容的可信度低,但其信息内容随着用户认知的变化、时间的不断更新和其他用户的影响,经过两次编辑而得到 answer_a2,其信息可信度值从 0.348 增加到 0.56。基于此可以看出构建数据起源模型来评估知乎信息的可信度是具备有效性的。

5 结语

本文从信息传播过程和用户两个角度对知乎平台的信息可信度进行评估,并构建可信度评估框架。采用数据起源的方法,根据所确定的知乎信息传播场景,抓取相关起源信息,建立 PROV 起源模型,采用 RDF 对其进行描述,通过计算传播路径上用户节点的可信度,从而得出知乎信息的可信度。本文研究成果为改善信息可信度评估方法以及优化社会化问答社区信息质量提供了一条新思路。此次研究为初步的尝试,在知乎信息选取的范围和类型上还是存在着一定的局限性,用户关系的复杂性和不同的话题类型都会给评估结果带来一定的差异。未来可将评估对象的数量和类型扩大,从而提高信息可信度评估结果的准确性。其

次,针对知乎用户的特点所制定的加权方案存在主观因素,后续可以考虑更多的影响因素来优化加权,从而进一步完善信息可信度评估结果。如何运用PROV模型对大量的起源信息进行显示以及管理,将成为未来研究的重点。

参考文献:

[1] 知乎:截至9月份,知乎个人注册用户总数超过了1亿[EB/OL].[2018－12－24].http://www.sohu.com/a/193351816_812860.

[2] CNNIC.第41次中国互联网络发展状况统计报[EB/OL].[2018－01－23].http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/P020180305409870339136.pdf.

[3] 曹高辉,胡紫祎,张煜轩,等.基于外部线索的社会化问答平台信息质量感知模型研究[J].情报科学,2016,34(11):122－128.

[4] 姜雯,许鑫.在线问答社区信息质量评价研究综述[J].现代图书情报技术,2014,30(6):41－50.

[5] 王平,程齐凯.网络信息可信度评估的研究进展及述评[J].信息资源管理学报,2013,3(01):46－52.

[6] 李保珍,王亚.社交媒体环境下网络信息可信度评估研究综述[J].情报学报,2015,34(12):1314－1321.

[7] JEON J, CROFT W B, LEE J H, et al. A framework to predict the quality of answers with non-textual features[C]//Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM Press, 2006:228－235.

[8] AGICHTEIN E, CASTILLO C, DONATO D, et al. Finding high-quality content in social media[C]//Proceedings of the 2008 international conference on Web search and Web data mining (WSDM'08). New York: ACM, 2008:183－194.

[9] SHAH C, POMERANTZ J. Evaluating and predicting answer quality in community QA[C]// International ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2010:411－418.

[10] TIAN Q, ZHANG P, LI B. Towards predicting the best answers in community-based question-answering services[EB/OL].[2018－09－15]. http//www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6096/6334.

[11] 来社安,蔡中民.基于相似度的问答社区问答质量评价方法[J].计算机应用与软件,2013,30(2):266－269.

[12] 王伟,冀宇强,王洪伟,等.中文问答社区答案质量的评价研究:以知乎为例[J].图书情报工作,2017,61(22):36－44.

[13] OH S, WORRALL A, YI Y J. Quality evaluation of health answers in Yahoo! Answers: a comparison between experts and users[J]. Proceedings of the American Society for Information Science & Technology, 2012, 48(1):1－3.

[14] 贾佳,宋恩梅,苏环.社会化问答平台的答案质量评估——以“知乎”、“百度知道”为例[J].信息资源管理学报,2013,3(2):19－28.

[15] FICHMAN P. A comparative assessment of answer quality on four question answering sites[J]. Journal of information science, 2011, 37(5):476－486.

[16] KIM S, OH J S, OH S. Best-answer selection criteria in a social Q&A site from the user-oriented relevance perspective[C]// Proceedings of the 70th annual meeting of American Society for Information Science and Technology (ASIST). Silver Spring: American Society for Information Science and Technology, 2007:1－15.

[17] 李晶.虚拟社区信息质量建模及感知差异性比较研究[D].武汉:武汉大学,2013.

[18] 孙晓宁,赵宇翔,朱庆华.基于SQA系统的社会化搜索答案质量评价指标构建[J].中国图书馆学报,2015,41(4):65－82.

[19] 施国良,陈旭,杜璐锋.社会化问答网站答案认可度的影响因素研究——以知乎为例[J].现代情报,2016,36(6):41－45.

[20] W3C. PROV-O: the PROV ontology[EB/OL].[2018－01－11]. http://www.w3.org/TR/2013/REC-prov-o20130430/.

[21] 闫浩.认真的人永远存在:关于知乎,这可能是最真诚的一篇分享了[EB/OL].[2018－03－28].https://www.huxiu.com/article/147187/1.html? f=index_feed_article.

[22] 倪静,孟宪学.关联数据环境下数据溯源描述语言的比较研究[J].现代图书情报技术,2013(2):18－23.

[23] 刘清松.中文微博信息可信度分析方法研究[D].北京:北京信息科技大学,2015.

[24] 知乎.什么是知乎的“优秀回答者”标识?[EB/OL].[2017－12－23].https://www.zhihu.com/question/48509984.

[25] 倪静,孟宪学.PROV数据溯源模型及Web应用[J].图书情报工作,2014,58(3):13－19.

作者贡献说明:

张婷:确定论文主题,负责论文研究框架设计、资料收集与整理、论文写作;

齐向华:负责论文研究内容的指导。

Credibility Evaluation and PROV Model of Zhihu Information

Zhang Ting Qi Xianghua

School of Economics and Management, Shanxi University, Taiyuan 030006

Abstract: [**Purpose/significance**] This paper aims to construct a PROV provenance model and user credibility evaluation index for information dissemination process, quantify the credibility of information, and enrich and improve the method of credibility evaluation of socialized Q&A community platform. [**Method/process**] The paper analyzed the credi-

bility of the data origination concept assessment information from the perspective of the information dissemination process, traced and recorded the source and dissemination of the information by establishing the relevant PROV data provenance model. The process, combined with the user credibility scores involved in the information dissemination process, was used to calculate the quantitative results of the credibility of the information. [**Result/conclusion**] Through the evaluation of the credibility of information, the information credibility evaluation method is further improved, which provides a new idea for optimizing the quality of community information.

Keywords: data provenance PROV-O Zhihu information credibility

关于举办 2019 年全国图书馆新型服务能力建设学术研讨会的通知

图书馆发展正在从资源驱动走向服务主导的新时代。图书馆发展的根本问题是服务问题,而服务的根本问题是服务能力。图书馆的服务能力需要以用户需求为牵引,从传统的以文献、馆舍和馆内服务,向嵌入式、深度的学科服务、情报服务、数据服务、出版服务、智库服务、智能服务和智慧服务等新型服务延伸和拓展,实现图书馆与图书馆服务的重大转型与变革,增强图书馆的服务功能与服务效果。

为此,现定于 2019 年 7 月 31 日至 8 月 3 日在黑龙江省齐齐哈尔市召开 2019 年全国图书馆新型服务能力建设学术研讨会。会议将邀请国内图书馆服务领域重要专家学者与一线工作者共同探讨和研究新形势下图书馆新型服务能力提升的理论与实践问题,总结和分享我国各级各类图书馆过去几年在新型服务方面所做的成功探索,分析和解决当前和未来图书馆新型服务能力建设面对的新问题与新挑战。

一、组织机构

主办单位:中国图书馆学会学术研究委员会

承办单位:齐齐哈尔医学院图书馆

《图书情报工作》杂志社

二、会议主题

图书馆新型服务能力建设

分主题:

1. 新时代图书馆的发展要求与转型趋势
2. 学科服务的发展模式与创新变革态势
3. 情报分析与竞争力研究与服务
4. 科学数据管理与服务理论与实践
5. 开放出版与图书馆出版服务
6. 图书馆的智库能力与智库服务
7. 智能图书馆与智慧图书馆服务
8. 新型图书馆服务体系建设与服务评价
9. 其他

三、举办时间及地点

时间:2019 年 7 月 31 日至 8 月 3 日(含报到和离会日期)

地点:黑龙江省齐齐哈尔市

四、参会人员

各图书馆馆长、分管副馆长、部门正副主任,业务骨干,以及从事图书馆服务研究的专家、学者、研究生;图书馆学情报学学术期刊主编、副主编、编辑;相关企业与技术人员。

五、会议报名

(一)费用

本次会议收取 850 元/人(交通、住宿费用自理),由齐齐哈尔医学院开具会议费发票。交费方式:

1. 微信缴费:7 月 15 日前将会议费通过微信缴费,报到现场领取发票,缴费流程如下:

(1)关注"齐齐哈尔医学院微服务"微信公众号

(2)点击页面下方左下角"在线缴费"→"其他缴费"栏目下"全国图书馆新型服务能力建设学术研讨会"→填写缴费信息(包括参会人员姓名、单位名称、纳税人识别号、联系电话等)→点击"确认提交"→微信支付(请根据报销要求选择支付卡并保留截图)



2. 银行汇款:7 月 15 日前将会议费汇至以下账户,报到现场领取发票。汇款时请注明"齐齐哈尔会议 + 单位名称 + 姓名"。转账需要通过单位公对公转账,个人转账仅能开个人抬头的发票。

开户行:中国银行齐齐哈尔分行

账　号:167702698126

收款单位:齐齐哈尔医学院

(二)报名方式

会议采取网上报名方式,请通过中国图书馆学会网站右上角"中国图书馆学会会员管理与服务平台"入口注册(如已有账户,则不必注册),以"个人用户"注册成功并登录后,在"我的首页"会议板块找到本次会议名称,点击"参加"提交报名申请。报名截止日期 2019 年 7 月 15 日。

(三)报名联系人

文丽:0452 - 2663511,13836279368,wenli67@126.com

徐明卉:0425 - 2663582,18604521966,395677258@qq.com

六、报到和住宿地点

酒店 1:齐齐哈尔万达嘉华酒店

费用:单人房 350 元/间(含单人早餐);标间 400 元/间(含双人早餐)

酒店 2:嫩江宾馆

费用:240 元/间(含双人早餐)

(注:会务组在以上两个酒店均设报到处)

中国图书馆学会学术研究委员会

2019 年 3 月 22 日